DOCUMENT RESUME

ED 233 049 TM 830 483

AUTHOR Wigdor, Alexandra K.; And Others
TITLE On Developing More Efficient and Imaginative Procedures for Conducting the National Assessment of Educational Progress. Report of the HumRRO Team.
INSTITUTION Human Resources Research Organization, Alexandria, Va.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO HumRRO-FR-PRD-82-1
PUB DATE Nov 82
CONTRACT 400-82-0016
NOTE 40p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cost Effectiveness; *Educational Assessment; Educational Research; Elementary Secondary Education; *Federal Programs; Planning; *Program Descriptions; *Program Effectiveness; *Program Improvement; Resource Allocation; Testing; Trend Analysis
IDENTIFIERS *Human Resources Research Organization; *National Assessment of Educational Progress

ABSTRACT

This report provides a blueprint for reorganizing the governing structure of the National Assessment of Educational Progress (NAEP) and suggests new procedures of test development and test administration that will increase the interpretability of the assessment while at the same time making more staff and monetary resources available for research and user services. Further, the plan includes provisions for protecting the trend analysis function of NAEP by assuring a satisfactory level of comparability between the new assessment and prior assessments. A number of fundamental changes are necessary if the objectives of greater efficiency and policy relevance are to be met. Among the recommended changes for NAEP are: (1) a full-time, professional staff for the Policy Assessment Committee; (2) the introduction of methods for developing objectives and items that are more cost efficient, provide for input from wider audiences, and result in a product that could be widely used by state and local school districts; (3) creation of a "national" item bank; (4) introduction of simplified methods of test administration; (5) revision of sampling procedures to eliminate cluster sampling; and (6) use of item domains. Primary type of information provided by the report: Program Description (Operating Policies); Procedures (Conceptual). (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Final
Report
82-1

HumRRO
FR-PRD-82-1

TM

HumRRO

ED233049

Report of the HumRRO Team

On Developing More Efficient and Imaginative Procedures for Conducting the National Assessment of Education Progress

Prepared by:

Alexandra K. Wigdor
Robert Sadacca
The Human Resources Research Organization

Richard K. Hill
Stuart R. Kahl
The RMC Research Corporation

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

November 1982

TM 830 483

Prepared for:

National Institute of Education
1200 19th Street, N.W.
Washington, D.C. 20208

Under

Contract No. 400-82-0016

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

2

TABLE OF CONTENTS

Introduction 1

Summary of Recommendations 1

The Assessment Policy Committee 3

Areas of Governance 3

Selection of APC Members 5

Relationship Between the APC and the Grantee 8

APC Working Procedures 9

The Assessment Policy Committee Staff 10

Constructing the Assessment 12

Development of Objectives 14

Development of Items and Test Packages 18

Administering the Assessment 23

Sampling Issues 23

Test Administration 25

Paced Tapes 28

Comparability Issues 28

Local Option 30

Reporting, Dissemination, Service and Research 32

Up-Grading User Services 33

Benefits to State and Local Agencies 33

Research 35

Reporting and Dissemination 37

INTRODUCTION

The following report is presented by HumRRO and its subcontractors, RMC Research Corporation and Intran, in fulfillment of contract No. NIE-400-82-0016. It provides a blueprint for reorganizing the governing structure of NAEP and suggests new procedures of test development and test administration that will increase the interpretability of the Assessment while at the same time making more staff and monetary resources available for research and user services. Further, the HumRRO plan includes provisions for protecting the trend analysis function of the National Assessment by assuring a satisfactory level of comparability between the new Assessment and prior assessments.

Summary of Recommendations

A basic principle of the HumRRO team's approach is conservancy. We have tried to build upon the first twelve years experience with the National Assessment, retaining those elements of proven value and learning from those practices that have proved to be less than satisfactory. We have concluded, however, that a number of fundamental changes are necessary if the objectives of greater efficiency and policy relevance are to be met. Chief among the changes recommended in the following pages are:

- The provision of a full-time, professional staff for the Policy Assessment Committee so that there will be continuity in the supervisory role accorded the APC by law. This is the mechanism that will empower the APC to exercise, for the first time, the leadership intended by Congress;

- The concentration of advisory functions in the Assessment Policy Committee, which will include within its own membership sufficient psychometric and statistical expertise to provide technical direction without having to rely on a competing entity for technical advice;

- The introduction of methods for developing objectives and items that are more cost efficient, provide for input from wider audiences, and result in a product that could be widely used by state and local school districts;

- Use of state assessment item pools as a source of items and creation of a "national" item bank from which states in turn could draw;

- Involvement of the same people from the beginning to the end of an assessment cycle, thus integrating the processes of developing objectives and items with analysis, and the reporting and dissemination of results;

- The introduction of simplified methods of test administration, particularly the use of local school personnel as administrators, that will result in great savings to NAEP, while involving more local officials in the NAEP process;

- Provision for a program of discretionary participation in the Assessment--a method by which states and local school districts can administer the NAEP tests to their students and receive truly comparative results at nominal cost;

- Revision of sampling procedures to eliminate cluster sampling--thus modestly reducing sample size while maintaining the same standard errors of estimate as currently obtained;

- Use of item domains as a strategy for item selection and the application of item response theory to scale the domains;

- The allocation of resources made available by more efficient development and administration to research and dissemination activities, including the provision of direct technical assistance to states and local school districts wishing to use NAEP materials and data. After a transition period, this would become a self-supporting unit within NAEP, operating on a cost-reimbursement basis paid for by the users; and

- The expansion of student and principal questionnaries and the institution of teacher questionnaires to enhance the explanatory or interpretive power of the assessment.

THE ASSESSMENT POLICY COMMITTEE

According to the provisions of P.L. 95-561, the educational organization selected by the National Institute of Education to carry out the National Assessment of Educational Progress is required to "delegate authority to design and supervise the conduct" of the assessment to an assessment Policy Committee (APC). The Grantee is instructed to establish the Assessment Policy Committee according to certain specifications set forth in the Act. In a very real sense, then, the Assessment Policy Committee is the creature of the Grantee; its membership is selected by and its authority derives from the educational organization that wins the competition to manage the National Assessment. At the same time, it is clearly the intent of Congress that the Assessment Policy Committee provide real leadership in structuring the Assessment, determining its goals, and in making it useful to educators and researchers.

In order to get the most out of this symbiotic relationship and to promote the leadership role envisioned by Congress for the Assessment Policy Committee, it is crucial that the APC's role be carefully defined and that it be provided with sufficient staff and the necessary instrumentalities to effect its policy decisions. There are two fundamental weaknesses in the existing governing structure of the National Assessment: The first is the lack of any effective mechanism of quality control from without or within. The administering agency does not have statutory authority to exercise such oversight, and the existing NAEP organization has not developed adequate internal mechanisms for the purpose. The second is the failure of the Assessment Policy Committee to fulfill its oversight role; like the British monarchy, it reigns, but does not rule. Both of these problems are addressed in our rethinking of the role and make-up of the Assessment Policy Committee and its relationship to the NAEP staff. We will suggest a number of structural changes that will strengthen the position of the APC, chief among them the creation of an independent core staff whose sole function is to serve the Committee. We will suggest certain new activities for the APC, such as an annual program review, that will promote a more robust appraisal of the Assessment by the APC and NAEP staff. Finally, we will suggest a number of ways in which staff members can support the APC and promote its effectiveness.

Areas of Governance

To carry out its statutory mandate, the Assessment Policy Committee must exercise governance in three broad areas: policy development; technical direction; and program administration.

A. Policy Development. The assessment Policy Committee will need to scrutinize the development and administration of policy within the broad framework established by statute and given definition by the National Institute of Education. The HumRRO analysis accepts as a given

the interest of the administering agency in maintaining data comparability to protect the trend analysis function of the National Assessment; in addition, it advances the agency's concern to increase the utility of assessment data to state and local authorities, federal policy makers, and educational researchers. With these two overarching goals in mind, the central task of the Assessment Policy Committee should be to redirect a significant portion of the available manpower and monetary resources to the currently relatively neglected areas of research, dissemination, and service. Specific suggestions for reducing the costs of test development and administration activities are described elsewhere in this document; the APC will need to monitor these and other initiatives to ensure that intermediate decisions actually do contribute to the larger goal of increasing the utility of the National Assessment without impairing its comparability with prior assessments.

B. Technical Direction. Technical issues will loom large during the period of transition. The Assessment Policy Committee must be prepared to oversee staff development of a new sample design, and the development of equating techniques with which to relate the new reference population to the old and the new test forms to the prior assessments. It must, in addition, supervise the ongoing process of test development, formulation of objectives, and the creation of linkages between test items and objectives.

As a first step, we propose that the Assessment Policy Committee include, in addition to the categories of people specified by statute, a measurement expert and a statistician in order to strengthen its role in providing technical direction for the Assessment staff. It is intended that these two experts, along with three other members of the Committee, form a Subcommittee on Measurement Issues. The special role of the Subcommittee will be to designate those technical issues most needing the Committee's attention, to gather, with the aid of staff, the information needed by the Committee to make sound decisions, and to marshal the best professional opinion where alternative strategies seem equally compelling.

The HumRRO team is convinced that the Assessment Policy Committee will function more effectively in providing technical direction if it contains within its own membership sufficient technical expertise to make the entire Committee conversant with basic psychometric and statistical requirements. This alteration of current practice, in which technical advice is the purview of a separate, and in some sense competing, advisory body, is recommended as a means of keeping psychometric and statistical considerations from overwhelming the process of policy formation. As a sizeable research literature shows, it is all too easy for non-specialists to be awed by the ex cathedra pronouncements of quantitative scientists. This is far less likely to be the case if the technical experts participate in the give-and-take of the committee process and present their advice within the context of the larger issues facing the Committee. The mechanism of a Subcommittee on Measurement Issues will provide the means of focusing on technical matters. To maintain an appropriate balance, we have recommended that

the two technical experts on the Subcommittee be joined by three non-specialists; further, we would bar the technical experts from holding the chair of the Subcommittee or the main Committee.

C. Program Administration. The third area of APC jurisdiction is administration, including fiscal management, program administration and dissemination activities. The APC, working closely with the senior staff of NAEP, will determine basic policy for the allocation of resources to the various programs, e.g., test development, user services, research.

A number of important new activities are recommended for the Assessment Policy Committee as part of its supervision of the NAEP programs. First, the APC staff, functioning as an internal auditor, should make periodic reports on each program for review by the APC. In addition, the APC, with the assistance of its staff, should conduct an annual review of the budget. We also strongly recommend that the Assessment Policy Committee institute indepth program reviews, to be carried out seriatim, one per year, to assess the quality and effectiveness of each major NAEP activity.

Our recommendation for one thorough-going program review each year holds out the best hope for making the Assessment Policy Committee ruler in fact as well as in name. If the APC has its own high-level staff, it will be guaranteed that measure of independence conducive to penetrating, critical inspection of the enterprise it serves. And it will have a task worthy of its serious consideration. It is likely that at least some of the annual program reviews will require the participation of outside experts, and provision will need to be made in the budget to support them. The amount would be modest, certainly commensurate with the benefit of having fresh perspectives brought to bear. The annual review would also seem to provide an appropriate avenue for the cooperative participation of the National Institute of Education in the National Assessment. Representatives of the agency could contribute the federal perspective on current educational needs and policy issues, which would be of real assistance to the APC as it weighs the content, direction, and effectiveness of particular program areas.

Selection of APC Members

The educational organization selected by the National Institute of Education to carry out the National Assessment has statutory responsibility to establish an Assessment Policy Committee composed of:

a. two representatives of business and industry

b. three members representing the general public

c. one chief state school officer

d. two state legislators

e. two school district superintendents

f. one chairman of a state board of education

g. one chairman of a local school board

h. one governor of a state

i. four classroom teachers

To that list we recommend the addition of:

j. one measurement expert

k. one statistician

The clear congressional intent in so specifying the membership of the Assessment Policy Committee was to ensure that all of the reservoirs of relevant experience be plumbed and all of the major participants in the educational enterprise be represented in the body empowered to design and supervise the National Assessment. That intent should guide the process of identification of potential candidates and the final selection of a balanced committee. In seeking appropriate representatives of business and industry, for example, the selection criteria should include familiarity with the skill requirements of entry-level jobs, employee selection procedures, job-training programs, or other sorts of experience that will provide insight into the intersection of formal education and the preparation of the young for the world of work. Similarly, the four classroom teachers should be chosen to represent as broadly as possible the experience of the entire profession. The first criterion should be that each is currently or has very recently been working in the classroom. Second, each should have had experience with group-administered standardized tests and should be familiar with the domain-referenced approach to educational assessment. Additional selection criteria would be aimed at creating an appropriate balance of geographical regions, size of school, socio-economic character of the community, grade level taught, area of subject-matter specialty, and special teaching experience such as bilingual education, teaching of advanced or highly creative students, or vocational education.

In addition to seeing that the Assessment Policy Committee includes in its membership the congressionally required state and local policy makers and practitioners, public members and experts, the selection plan should emphasize a number of other criteria. The first of these is excellence. Each member of the Committee should be recognized as outstanding in his or her field of endeavor. In some cases, that will mean academic or scholarly excellence; in some, excellence will mean the possession of high-level skills as demonstrated by teaching awards or other expressions of peer approval; in the case of the public members, excellence will be found in service to education, community activities, and other areas of public involvement.

A second criterion is relevant professional experience. Each member should bring to the Committee either substantive expertise or practical experience of some aspect of educational policy-making or practice. It is particularly important that the group as a whole exhibit a variety of experience and knowledge as an impetus to flexible and imaginative policy making. Certainly a good number of the members must be accustomed to considering policy questions and conversant with the issues that fuel public debate over education.

Another equally important consideration is the ability to work well in a group. Every committee requires time to learn to work together effectively, to develop a working style, and to reach a consensus on basic goals. Good staff support can aid the process immeasurably, but the success of a group depends ultimately upon selecting as members people who have the capacity to listen to others as well as to speak, who have the self-discipline to concentrate on the task at hand and resist attractive digressions, and whose basic attitude toward others is one of respect rather than challenge. The dynamics of the committee process are such that brilliance tends not to be exploited in the absence of these other interactional skills.

A final principle of this selection plan is that the Assessment Policy Committee should be a national committee; that is, a distillation of the diversity of the American people and the variety of its educational systems. Although each candidate should be vetted initially for excellence and appropriateness of professional experience and group-process skills, the selection process must be consciously designed to locate candidates from all geographic regions and from major racial, ethnic, and gender groups. The simultaneous application of all these criteria will result in a balanced committee which will command the confidence of the educational community, local, state, and national officials, and the general public.

The process of assembling a list of qualified candidates, if it is to be more than a hit-or-miss affair, must be a structured, systematic search. As a first step, the Grantee should meet with program officers of the National Institute of Education to discuss possible candidates and to set up a liaison group which will serve as a conduit of information and advice between the agency and the Grantee during the establishment of the APC and, indeed, for the life of the contract. At this time the Grantee should also establish contact with present and former members of the APC and other NAEP advisory committees to discuss the functioning of the APC, strengths and weaknesses of the committee process at NAEP up to this time, the qualities that appeared to them most useful for the particular tasks that faced them as committee members, and their suggestions for membership on the newly formed Committee.

The identification of potential committee members will involve contacting leading scholars, educators, directors of educational associations and professional groups, and political staff members to discuss proposed candidates and receive additional suggestions. A survey of the programs of recent annual meetings of teachers groups, school board associations, and other relevant organizations will rapidly

reveal the names of educators and public officials who are professionally active and interested in larger education issues, but who might not show up on the more traditional information networks.

In order to provide structure to this process and to allow eventual weighting of the qualifications of the candidates, it is important to explain the selection criteria to the people contacted for information and to take note of the judgments made about a candidate in each assessment category. This is a simplified version of the structured interview developed by personnel psychologists to lend greater validity and reliability to that universally used and notoriously erratic selection device. As the preliminary selection stage proceeds, the interviewer will see a gradual convergence of opinion about proposed candidates from which a "long list" of qualified candidates will emerge.

To further narrow down the list of candidates in each of the occupational categories specified by law, the next step should be a thorough review of the scholarly, professional, and/or public lives of the candidates. This will include gathering biographical data on each of them, reading a few recent articles authored by those who are active researchers or scholars, doing some research on the legislative or administrative record of those who are in public life. One would want to choose state legislators, for example, who exhibited a strong interest in educational issues during their years in office, perhaps by having introduced legislation to establish a state assessment or minimum competency testing program.

Having identified the two or three strongest candidates in each category, the Grantee can now move on to the finer distinctions that will make for balance in the committee as a whole. These will include balancing geographical areas, racial and ethnic identity, age and gender, and so on. It will be important to choose people with experience of large, urban situations as well as those familiar with small-town and rural America; people who know a good deal about educational theory and people who are in touch with political reality, so that the state-of-the-art and the art of the possible converge; people who will be in a position to devote a great deal of time to APC matters in addition to those whose other commitments will limit their participation.

Relationship Between the APC and the Grantee

The relationship between the APC and the Grantee is, above all, one of interdependence. The APC is established by the Grantee; its authority is delegated to it by the Grantee. Yet that authority is great. The Grantee is instructed by law to look to the Assessment Policy Committee for policy direction in the design and conduct of the National Assessment.

The collaboration between the two entities will be more or less productive depending on the resources devoted to making it work. It is our strong opinion that the Assessment Policy Committee can be of optimum benefit to NAEP only if there is continuity in the relationship.

If the role of the Committee is limited to coming to meetings three or four times a year, there to work through a large agenda book of rapidly skimmed and half digested action items, then its existence will be largely a formality. To have a significant substantive impact on NAEP operations, the APC must be enabled to bring the sustained attention of its members to bear on NAEP matters. The best way to achieve this end, we would suggest, is to provide the Assessment Policy Committee with its own permanent professional staff. This will allow the Assessment Policy Committee to develop a level of expertise about the National Assessment that is now denied it.

The APC staff would prepare reports on educational research, interim program reviews, issue briefs, position papers, and other documents designed to inform the Committee members and help them articulate policy questions. In the period between meetings, the APC staff would keep the Committee members up-to-date on the working of the NAEP programs, the most important issues and problems facing each section head, and the status of new initiatives and special projects so that they could come to meetings fully prepared to act.

Earlier in this discussion, we described the function of the APC staff as that of providing the Committee with an ongoing internal audit of NAEP operations. This will make an important contribution to quality control. Moreover, from the perspective of the Grantee, the existence of an APC staff will greatly increase the responsiveness of the supervisory body. At the request of their colleagues on the larger NAEP staff, APC staff members could confer with individual committee members about special issues within their area of expertise or poll the whole Committee so that NAEP staff leaders can rapidly get a sense of the Committee on pressing policy questions.

A professional APC staff will, therefore, promote more effective leadership of the National Assessment. It will make for a better informed Assessment Policy Committee, it will provide the means of more frequent and substantive communication between the Grantee and the Committee, and it will provide a mechanism for increased quality control.

APC Working Procedures

Because of the larger role envisioned for the Assessment Policy Committee under the HumRRO plan, it will need to meet on a quarterly basis. During the transition period, the APC should hold two meetings, the first of them to be held within 60 days of the awarding of the contract. During the first two-day meeting, the APC will need to familiarize itself with the procedures used by the Education Commission of the States in its administration of the National Assessment and study in detail the plan of operation proposed by the Grantee, including the equating techniques and other measures designed to provide continuity in the meaning of the data and in the technical services offered to states and localities.

In the period between the first and second meeting, the APC staff, in close collaboration with the Assessment staff, will work with each member of the Assessment Policy Committee to make sure that they are fully informed about the statistical, psychometric, and financial considerations that have led to the Grantee's proposed alterations in the sample design, assessment procedures, and dissemination activities. As a consequence, the APC will be prepared, at its second meeting, to formally approve, with whatever modifications of the Grantee's basic plan the Committee has deemed it necessary to introduce, a blueprint for carrying out the Assessment, including specific plans for using the materials that ECS will have developed for the 1983-1984 Writing Assessment and the 1984-1985 Reading Assessment, and for otherwise making full use of the experience gained since NAEP became operational in 1969.

At the completion of the transition period, the Assessment Policy Committee will begin a regular sequence of four annual two-day meetings, one devoted to review of the assessments under development; one to an annual budget review and discussion of special projects; one to research and dissemination activities; and one to a major annual review of one program area. It will also, with the assistance of its staff, supervise the formulation of a framework of objectives and the development of tests on an ongoing basis. This is an ambitious schedule of activities that will provide dynamic leadership for the National Assessment.

The Assessment Policy Committee Staff

The Assessment Policy Committee will exercise governance on a day-to-day basis through the APC staff. The APC staff will serve as a conduit of information from NAEP to the Committee and will communicate policy directives from the Committee to the Grantee. At the request of the Committee, its staff will prepare briefing documents, secure the advice of subject matter specialists on issues facing the Committee, formulate, in concert with their NAEP colleagues, assessment objectives for committee discussion, and otherwise support the Committee on a continuing basis.

To be adequate to the task, the APC staff should consist of a staff director; two senior research associates who will provide briefing documents on assessment issues, educational research, and educational policy questions to inform the APC decision-making process; an administrative assistant, who will serve as meetings coordinator and provide day-to-day financial management; and a secretary.

The staff director, working closely with the chairman of the Assessment Policy Committee, will organize the annual schedule of reports to the APC, plan the program and budget reviews and the quarterly meeting of the Committee, and supervise the financial management of the Committee budget. The staff director will meet regularly with the heads of the other major divisions of NAEP to discuss

policy and program activities. These senior staff meetings will be an important means of communication between the Assessment Policy Committee and the Grantee. They will enable the APC staff director to give the Committee early warning on emerging problems or to inform division heads of information about their program area that the APC has requested or is likely to need in the coming months. Above all, they will be a means of seeing that APC policy decisions are reflected in the operation of the Assessment programs.

Although cooperation between APC and NAEP staff members is vital, it is also imperative that the APC staff be independent of the regular Grantee staff. It must be clear in the minds of the APC staff members that they are working for the APC and not the Grantee, otherwise the critical review and independent appraisal functions of the APC will suffer. It is therefore highly recommended that either a separate subcontractor be selected to support the APC, or if the Grantee is sufficiently large, a separate major division of the Grantee be given responsibility for supporting the APC.

CONSTRUCTING THE ASSESSMENT

Standardized testing and assessment represent an attempt to evaluate attainment in a uniform, rational, and systematic manner. At the heart of the enterprise is the question of meaning. What is measured by a particular item type? What does it mean if a person or a cohort can or cannot answer test questions accurately? What is the relationship between test performance and ability or test performance and attainment in every-day life? One of the most difficult and challenging tasks for the test developer is to assign meaning to test performance.

There are many aspects to the task. The very process of standardization provides an important aid to interpretation: It allows the user of test data to infer differences in attainment levels between individual test takers or between successive cohorts of test takers as reflected in different score levels by assuring a high degree of uniformity in the conditions under which the test is administered as well as comparability of successive test forms. In other words, standardization significantly reduces the influence of extraneous factors on test performance so that the measure produced has some sort of objective reality.

The most important strategy for making test performance meaningful has been to introduce a comparative structure, for example, the comparison of an individual's performance with that of a norm group, the comparison of test performance across time, or the comparison of individual or group performance against some standard. Since its inception, the National Assessment has eschewed the first tactic, and will continue to do so. It is not designed to demonstrate individual differences within an assessment generation and does not, therefore, use the norm-referencing techniques typical of such testing programs as the Scholastic Aptitude Test, the American College Testing Program, and other tests designed to aid selection or placement decisions. The National Assessment has, however, made use of the other two tactics, that is, the comparison of age cohorts across time and, rather less successfully, comparison against some standard of performance.

Enough time has elapsed since the introduction of the National Assessment to generate a good deal of interest in its trend analysis function. This success and the present opportunity to rethink the overall design of NAEP have led a number of commentators to suggest annual assessments in the major skill areas. The idea does have certain attractions. It would, no doubt, increase public awareness of the National Assessment. In addition, annual reporting of each assessment area would allow the smoothing of sample results by computing moving averages across years of assessments. Nevertheless, we have concluded that the disadvantages outweigh the benefits.

We feel that annual assessments in the major skill areas would lead to a dilution of effort in all areas. One of the most likely things to suffer would be the support and participation of interested groups. With reading assessed once every five years, for example, the International Reading Association has been willing to focus its attention on the assessment and put its full weight behind dissemination of the results. It is not at all clear that the same level of commitment would be forthcoming once a year.

The internal process of planning and development would also be diminished, given a static funding situation and a much increased development load. Because the basic thrust of the HumRRO team proposal is to reduce development costs and redirect resources to research and evaluation activities, the effects of a move to annual assessments would be doubly debilitating.

A final and convincing argument against annual assessments is that national averages change slowly. Given the length of time it takes for national averages to change meaningfully, it seems reasonable to question whether annual assessments would provide data of sufficient value to justify the cost. Current standard errors are approximately one percent, which means that, using current sample sizes, an item would have to change its p-value by almost 3 percent from one year to the next to be interpreted as a statistically significant change. If more forms were developed while the total sample size remained constant, standard errors would be even larger. But how many items could be expected to change by 3 percent or more in a year? Even if the argument is extended to include tests as a whole, rather than just items, standard errors still would be larger than the expected real changes. The net result would be a series of small shifts, up and down, that would be continuously over-interpreted by special interest groups. Every small turn downward (whether due to sampling error or not) would be leapt upon by critics of current education; every small rise would be embraced by supporters. The net result would be several years of crying "Wolf" until no one paid much attention to the annual fluctuations. For followers of the California data, it is almost amusing to note the great press flurry that has attended the announcement that twelfth grade results declined this year, compared to last year's results, by one-tenth of one percent! In light of the damage that can accompany such pronouncements, however, the situation loses its humor. It would be better to allow a period of time in which significant change can manifest itself--perhaps five years is an ideal spacing--before conducting a second round of assessment. However, a compromise of every three years may avoid the above pitfalls while providing some of the advantages of more frequent assessments.

All in all, it seems better to leave the most successful interpretive element of NAEP essentially as it is (elsewhere in this report mechanisms for maintaining the comparability of assessments are discussed), and to focus upon strengthening the third tactic for investing test performance with meaning, i.e., relating test performance to a standard. In this area, the current administration of NAEP has

been weak to the point of raising questions about the continued existence of the Assessment.

It has been the practice of NAEP to try to invest its assessments with meaning through a two-step process of development: first, the development of assessment objectives by committees made up of educators and researchers; and second, the development by committee of test items that relate to each objective or subobjective. The design was consciously aimed at producing objectives and tests through a process of building consensus, and meaning was derived in good part from the apparent reasonableness (face validity) of the objectives and items. NAEP did not make use of sophisticated psychometric techniques to develop test items and relate them to one another, so it has not been possible to build up a structure of inference based on the psychometric properties of NAEP items.

The politics of consensus have not in this instance made for robust interpretation. Indeed, so cautious was the original NAEP policy that reporting was done at the exercise level and it was left to the user of NAEP data or reader of NAEP reports to decide how to interpret each individual item and whether a change in the percentage of students in the sample who performed the task correctly is meaningful. The reporting of test results by individual item is to assessment what the writing of chronicles is to history: the entire burden of interpretation is placed on the reader, who has no way of knowing whether he has sufficient information to draw conclusions, no grounds beyond reason and common sense for weighing the importance of the discrete bits of information, no way of judging whether or how the events chronicled are related to anything important in the world at large.

In the following sections, we will suggest a number of mechanisms that will anchor the assessment more firmly in interpretive structures so that readers of NAEP reports and users of NAEP data are given greater guidance in drawing meaning from assessment data. The task of drawing practical and policy implications from assessment data will continue to be difficult. One should beware of statistical models that promise instant meaning. At the same time, the application and continual evaluation of new psychometric methods will bring to the National Assessment the focused attention and energy of the research community, which must increase the likelihood of developing over time a useful and policy relevant interpretive foundation for NAEP data.

Development of Objectives

The HumRRO strategy for the development of assessment objectives takes cognizance of the goals of cost effectiveness and assessment utility. To achieve these goals we recommend three changes in the overall assessment design that will alter somewhat the function of the objectives in the assessment:

- The integration of objectives development and item development into a single process by maintaining the continuity of staff and consultant involvement throughout the entire assessment cycle, and through psychometric analysis of items.

- The application of psychometric methodologies to test development to provide an interpretive structure; this will lighten the interpretive burden that the objectives now are asked (and consistently fail) to bear.

- Maximum use of existing materials, particularly state assessment objectives and items.

The impact of these changes would be to reduce slightly the theoretical importance of the objectives; but by making them more consistent with state objectives, strengthening the links between objectives and items through psychometric analysis, and increasing the knowledge of developers by using them through an entire assessment cycle, the practical importance of the objectives would be enhanced.

In moving away from the consensus-based approach to the development of assessment objectives, and toward a more objective procedure, we have found the literature on organizational frameworks instructive. References on test construction have long advocated the initial listing of instructional objectives, followed by the development of items addressing each of the objectives. Unfortunately, objectives lists have usually taken the form of content outlines representing a single content dimension. In the past, NAEP objectives have been organized in this manner, with one notable exception. The 1977-1978 national assessment in mathematics used a content-by-process matrix in which the process dimension represented revised levels of Bloom's Taxonomy. The purpose of adding a dimension to an objectives framework is to assure adequate coverage of important domains. This also produces useful reporting categories. The reporting of "process" categories in the NAEP mathematics assessment produced information of great interest to many groups and provided an empirical basis justifying a widely felt need to shift curricular emphasis--namely a shift from a primary emphasis upon mathematical skills to an emphasis on problem solving capability as the ultimate goal of mathematical instruction.

The NAEP mathematics assessment illustrates a problem associated with the use of a development matrix. During the development stages, insufficient effort was given to making sure that items adequately covered the process levels and thus the "cells" of the matrix. It was not until instruments were being administered that a very sparse coverage of mathematical "understanding" was revealed.

There are no "best" dimensions to use in a development framework. While the 1982 and 1983 Connecticut assessment in social studies did use Bloom's taxonomic levels, it also found two content-related dimensions useful: (1) discipline (history, geography, sociology, political science, etc.); and (2) theme (differences and similarities among

peoples, rights, duties, responsibilities; interdependence and interaction; adaptation and change; causes, effects and resolution of conflict). The primary value of the approach is that it emphasizes the fact that there are many dimensions to knowledge and to ability that can be objectively measured, and that a carefully designed assessment can be constructed so as to produce information on a large number of dimensions. The application of psychometric techniques to the development of tests and objectives will help the development team tease out the kinds of skills and knowledge elicited by various item types and thereby refine and enrich the power of the assessment to measure dimensions of interest.

Procedures such as these can result in a national assessment of great utility to many different kinds of users. But a development framework well suited to the purposes of one group--say, state and local school personnel--will not necessarily meet the needs of other users. While one would expect that the majority of objectives and items produced to "fill" one framework would be usable in a different framework as well, there is little question that use of the second framework would lead to the inclusion of objectives and items which otherwise would be overlooked. For example, the quality of the products of secondary investigations of NAEP data in the past has been marginal because NAEP assessments were not designed to accommodate these other purposes. In other words, secondary analyses of NAEP data in the past have suffered because the NAEP data was not intended to be used to answer the questions addressed by the secondary analyses. The only way this problem can be avoided is by taking these other possible uses of NAEP data into consideration during the development stages of an assessment.

While NAEP traditionally has involved a variety of consultants in development activities, decisions regarding the objectives to be assessed are said to have been left to a handful of university educators. Furthermore, those individuals have been somewhat restricted as a result of policies established early in NAEP's history when attitudes toward many assessment characteristics were very different from what they are today.

Whereas in the past, the participation of many groups in development activities has often been superficial, it seems particularly important now that the various users of NAEP be systematically involved in the development of objectives and test items. State and local content area specialists as well as state assessment personnel have been generally neglected in the development process. This seems unwise considering that the states are intended to be primary beneficiaries of NAEP services, materials, and results. NAEP could benefit tremendously from a closer working relationship with the states. Our plan for a more active Assessment Policy Committee will bring greater state and local influence to the overall design of the assessment. Our recommendation that state objectives and item pools be used will increase the contribution of the states to the content of the assessment. In addition, the active participation of state personnel in the nuts-and-bolts decision-making regarding assessment objectives is desirable, and we would suggest that they are used as consultants and reviewers.

The advantages of incorporating various research perspectives into the assessment have already been discussed. Such perspectives would lead the development team to incorporate many more variables than in the past. It is also important to incorporate the priorities of various funding agencies (e.g., NIE and NSF) into the overall planning of an assessment in any particular content area. There is no reason, with the development of additional demographic and other survey questions addressed to special issues, with well planned item development, and with effective item assignment to packages, that the regular assessments cannot serve several purposes. Teacher and principal questionnaires can be used to collect far more information than has previously been collected using NAEP's usual principal questionnaires. Information could be gathered at little additional cost on program characteristics, usage of materials, classroom practices and teacher training and experience much like the information gathered in the NSF-funded study, the National Survey of Science, Mathematics, and Social Studies Education (Weiss, 1978), conducted by the Research Triangle Institute.

In short, the development approach proposed to make NAEP data more useful calls for substantive involvement of the potential "users" in the development. NAEP utility would be further assured because the new NAEP design would maintain continuity of staff/consultant involvement throughout an assessment in a particular area. In the present system, once the development of an assessment is completed, the coordinators move on to the development of assessments in other areas. In many instances, data from the first assessment are analyzed and interpreted by persons with no involvement in the development process and no experience in the content area. This approach was questionable when NAEP restricted itself to the mere reporting of findings, but is clearly unacceptable if NAEP is to adopt a broader role. Those involved in the development of an assessment must be heavily involved in subsequent activities associated with that assessment. Such activities would include analysis; reporting; preparation of manuscripts for journals, presentations, and workshops; and consultation with, and training sessions for, users of NAEP data. NAEP's current approach to user services primarily employs persons not involved in the actual conduct of the assessment.

In consideration of the above discussion, the HumRRO team suggests revised procedures for the development of objectives. Assessment development activities will require at least three content area specialists--one with expertise in language arts education, an expert in math/science education, and a social studies/citizenship expert. These people will coordinate the development activities within their respective areas. Whether an additional coordinator would have to be engaged for other assessment areas such as art, music and literature could be decided when the basic system is in place. The content area experts will function throughout an entire assessment cycle--i.e., they will not only coordinate development activities, they will also be heavily involved in dissemination, technical assistance and research activities.

The first task of the content area specialists will be a careful review of the existing objectives for each assessment area. Some of these have been fine-honed over the years and enjoy a good reputation; others have been less than successful. Next will be the collection of existing materials from the many states which have already put considerable time and effort into the development of objectives and test items. RMC's experience in the past has shown that the states are generally quite willing to share their materials and, in fact, flattered to be able to do so. Equally valuable sources of existing materials which will be useful during the development of objectives are the professional organizations of teachers and teacher educators in the various content areas.

Using the existing materials as a starting point, the assessment team leader would then coordinate consultant activities, using a slightly different approach from that of ECS. Only one or two groups of consultants would meet to produce a proposed objectives framework and set of objectives. Part of this assessment planning will be the consideration of broader goals of the assessment--for example, the research questions to be addressed and the implications for test development. These consultant groups would consist of six to eight members representing teachers, university educators, researchers, experts in the appropriate discipline, and state department of education content specialists. Appropriate professional groups might be asked to nominate candidates.

With the direction of the Assessment Policy Committee, the development committees would establish a proposed set of objectives and decide upon the knowledge and skill dimensions and other variables to be measured. This plan would be reviewed by a much larger number of individuals representing the same categories of NAEP users. The review process would be accomplished by mailing materials to the consultants for their analysis and comment. RMC has had a great deal of success using the mail-review process. It is considerably more economical than conducting a large number of consultant meetings across the country, and often more efficient in terms of focused attention to the task. On the basis of this process of consultation with and advice from NAEP users, the Assessment Policy Committee would produce the final set of objectives to guide the process of test development.

Development of Items and Test Packages

In the matter of test development, we suggest two major revisions of current practice, one which will significantly decrease the cost of item development, and a second which, through statistical scaling, will allow the creation of item sets related to knowledge or skill domains and representing a specified range of difficulty levels.

As part of its user services and technical assistance program, NAEP has for years made its methodologies and some test items available to state assessment projects. This has been a generally useful practice,

although the states have nowhere near the fiscal resources necessary to duplicate the highly refined NAEP test administration procedures and have often found that NAEP items do not address objectives deemed important by the states. Given the developments of the last decade in the states, however, we suggest strongly that NAEP has remained for too long in a posture of assuming that the only way for information to flow is from NAEP to the states.

The states, despite their limited resources, have moved well past NAEP in some areas--most notably in item development. Many states have large pools of items that are of good quality--often superior to the items coming from NAEP. With this wealth of useful material available from the states, it no longer makes sense for NAEP to initiate a new development effort for every assessment cycle. Therefore, an essential recommendation of the HumRRO team is that NAEP draw from the item pools developed for state assessment programs, and not develop new exercises before making a thorough review of this vast supply of high-quality materials available from state education agencies.

As evidence of the feasibility of this recommendation, we offer the case of HumRRO subcontractor, RMC Research Corporation, which recently completed the development of a 1,000-item bank for Massachusetts, matching each item to a specific objective and item type. The entire pool was scaled on a sample of 12,000 students, and 6 equivalent forms of reading tests (55 items per form) and math tests (59 items per form) were developed from the bank. The total cost for this project was under $75,000. One reason for this low cost was the substantial use of items developed by other states. For assessing reading levels, RMC developed fewer than 100 new items; all other items were supplied free of charge by assessment programs in other states. Such a method of developing a pool of items costs far less than current NAEP practices, which basically ignore the plethora of high-quality materials, especially in reading, writing, and mathematics, available at no cost. By building on these free materials, NAEP would be able to maintain high-quality item pools even if inflation drives development costs higher.

There are several distinct advantages to the maintenance of a large item pool. First, of course, is cost effectiveness. Second, while a limited number of items may be kept secure for use in subsequent assessments to monitor changes in national performance, the vast majority of the exercises would be available for states to utilize, thus giving the states a much superior set of items from which to choose than they could possibly develop on their own or borrow from the released exercise sets currently available from NAEP. Moreover, this "national item bank" would in no way raise the specter of national standards or a Federal curriculum since the source of most of the items would be the states themselves and because the pool would be large enough for each state to tailor its test package to its own objectives. Third, any individual state would, over the years, find that NAEP could supply national data on a number of items it had selected for its own assessment. Indeed, this seems to be one of the most fruitful ways to make NAEP useful to the states, and, if our suggestions for simplifying NAEP administration procedures to bring them more in line with state

practices are implemented, current questions about the validity of comparisons the states make with national levels of performance on test items will be considerably eased.

There will undoubtedly be domains not currently assessed by NAEP that the Assessment Policy Committee will decide ought to be the object of special studies. One thinks immediately of computer literacy. In such cases the NAEP staff may well have to take a more active role in item development. But this will be an important opportunity to offer the states, through their use of the NAEP item pool, the chance to work together in assessing new educational needs or in evaluating the effects of new teaching strategies on test performance.

The creation of a national item pool comprised of questions drawn from state assessment programs can be easily and rapidly accomplished. And it should bring great improvement in the utility of the assessment to the states in a relatively short period of time. This recommendation has the benefit of being easy to effect, of being not only cost efficient, but of offering an absolute reduction in costs, and of promoting a much closer, cooperative relationship between NAEP and the states.

Our second recommendation with regard to test development is more complicated, less likely to produce immediate benefits, but important to the long-term, research-based interpretation of Assessment results. Rather than devoting so much energy to a consensus-building process of objective and item development in order to try to give the assessments meaning--an approach that has proved wanting--we suggest streamlining the committee system and focusing attention instead on the application of psychometric methodologies to the problems of developing relationships between items, building evidence that items measure what they are intended to measure and whether they measure the same thing in different subpopulations, and ordering items by difficulty level.

Some of these methodologies are well established, while others, like item response theory, are in the early stages of practical application, and will only show their full potential as NAEP (and outside researchers) refine the procedures on which their success depends.

Item response theory (IRT) is a promising enough approach that we think it should be introduced, perhaps incrementally, to supplement the present unsatisfactory system of interpreting results item by item, with averages of p-values (the proportion of students in the country answering a question correctly) being computed for certain clusters. Using IRT, items would be chosen to reflect the concept of a unified cluster of skills, rather than being discrete, individual items presented in groups.

IRT analysis is based on the ordering of items within a content domain by difficulty level. The application of IRT analysis to NAEP is attractive because it will allow objective demonstration of mastery levels from which inferences can be drawn about the nature of the

cognitive processes that underlie achievement at each difficulty level. It promises a real advance in understanding attainment and predicting numbers of people falling within defined levels of competency.

While IRT is a new methodology, it is not entirely untried. Several states, most notably California, have made use of IRT methods in order to produce linear data that permit revision of tests from year to year without revision of their scales or reporting categories. They have also found the methodology attractive because it allows easy interpretation of results at the local school level. And, because scales remain constant even when items change, longitudinal comparisons are simplified.

It may well be more difficult to define relevant and coherent content domains that lend themselves to unidimensional scaling for a national assessment than for a state assessment, and the process of criterion-referencing the domain scales would be difficult in any circumstance. For these reasons, it seems reasonable to recommend the gradual introduction of IRT methods beginning with a small number of item subsets, and gradually expanding as new content domains are defined and as research begins to support the interpretive value of the approach.

Our suggested general procedures of item and test development are consistent with those described in the previous section on the development of objectives, the rationale for those procedures applying equally as well to item selection and development. Under the overall direction of the Assessment Policy Committee, content area experts would be responsible for pulling together the item pools and coordinating the other development activities. The same consultant committee(s) would be involved, and the same mail-review process would be used. While substantial development of new cognitive items should not be required of the consultants, considerable effort will be required to develop instruments to measure variables previously not addressed by NAEP. Student and principal questionnaires could be expanded and the use of teacher questionnaries instituted to enhance the explanatory or interpretive power of the assessment.

In the experience of HumRRO and its subcontractors, field testing is an essential step in test development. It can point up problems with particular items--confusing or ambiguous wording, inadequate distractors, or simple typographical errors--and is important to the process of establishing item difficulty. If the Assessment is developed with an eye to the psychometric properties of the item sets, field testing will become even more important.

In the past, RMC has found interviewing a few students who have completed the field tests to be a valuable additional source of information--information, in many instances equally if not more valuable than that obtained from statistical analysis of field test data. This practice would be a useful complement to the NAEP field tests.

Traditional statistical item analyses should be performed on field test data, along with statistical techniques for detecting item bias. Of course, bias review should also be an assigned task of the APC and the teams involved in the item selection and development activities.

In addition to items, assessment procedures should also be field tested. This will assure that the administration manuals (which will be used by the teachers who will be the test administrators) are written clearly and are easily followed. Such field testing would be one of many quality control procedures necessary to guarantee uniformity of administration procedures.

The HumRRO plan envisions new assessment booklets, physically different from the old ECS packages. Since paced audiotapes would not be used (see the section on administration procedures below), many more items can be placed on a single page. It is quite conceivable that an expanded student questionnarie and all the test items assigned to a particular package would fit on 2 or 3 signatures (8 or 12 pages) thereby cutting printing costs considerably. (Over 50 pages were required in the past.) Such a format is more consistent with that used in state assessment programs. At the individual item level, the use of the "I don't know" option should be re-evaluated. States tend to not use this option even in items borrowed from NAEP. This practice further jeopardizes the validity of state-nation comparisons made in the past. If the use of "I don't know" is eventually discontinued, it should be phased out in such a way that data comparability over time is not sacrificed.

ADMINISTERING THE ASSESSMENT

For years, NAEP has been thought to represent the state of the art in large-scale assessment. In terms of sophistication, NAEP has indeed been a leader. However, NAEP methodologies were established in the context of relatively high funding levels. This was never the case for individual states; they have relied upon more cost-efficient assessment approaches using sampling and administration procedures very different from those of NAEP. (Many such differences are largely ignored when states compare the performance of their students on NAEP items to national averages.)

It is understandable that, faced with the drastic budget cuts of recent years, an organization would cut down on the materials and services it provides in order to preserve an existing effective methodology. In view of the changes NIE and the educational community in general would like to see in NAEP, however, we contend that alternative methodologies would be more appropriate for NAEP and that more can be done for less.

Sampling Issues

NAEP currently uses a multi-stage, stratified, age-specific, probability sampling plan to identify the students to be tested. The stages involve the division of the United States into 1,180 Primary Sampling Units (PSUs), the selection of sample PSUs, the selection of schools to be sampled within each selected PSU (in large population PSUs, this is a two-stage process) and, finally, the sampling of students at the correct age within each selected school.

The NAEP sampling plan is driven by several procedural constraints, most notable those deriving from a very complicated scheme of test administration. For example, all testing is conducted under the supervision of an official trained and sent into the schools by NAEP. The current method of sampling PSUs was designed in part to control the costs of employing these outside administrators; test sites are clustered so that each district supervisor need travel over a relatively small distance in going from school to school.

While the current sampling plan is reasonable given the constraints introduced by ECS's test administration procedures, major improvements in the sampling design--as well as significant cost savings--could be achieved by the simple device of using tests that can be administered without direct project supervision. First, there is the obvious saving of hundreds of thousands of dollars of expenses for district supervisors. Less obviously, but also important, would be the efficiencies accruing from eliminating the current school district PSUs as the initial selection stage. Gross clustering of this nature drives

up standard errors; elimination of such a practice would allow for the testing of fewer students, while still obtaining similar standard errors of estimate. The principal reason for using the current PSUs is to ensure that selected schools are near enough to each other to permit all assessment to be supervised by a project staff member. Elimination of these supervisors would permit desirable flexibility in the sampling plan.[1]

Another change that would promote interpretability as well as efficiency in the sampling plan is to sample students on the basis of grade, not age, and to make schools, rather than students, the final randomly drawn sampling unit. Since curriculum objectives are generally set by grade and not age, tracking students' educational progress through three grades (Grades 4, 8, and 11) would provide more relevant information to school administrators and other potential users of the data. Not only would a school-based sample be more relevant, it would make the assessment far less intrusive. The current practice of randomly selecting only a few students from each of several classrooms is complicated logistically and unnecessarily intrusive. Disruption of school routine would be minimized by testing all classes at the appropriate grade for one class period.

There are many positive aspects of the current sampling plan that should be continued. These include the use of multiple forms within each school, the eliciting of a high degree of school cooperation by using a variety of avenues to contact the schools, the use of stratification of schools to minimize sampling error, and the use of data available from state departments of education to minimize new data collection requirements.

However, by eliminating on-site administrators and paced tapes, revising age-level sampling to grade-level sampling, adding matrix sampling through the use of class packs (shrink-wrapped packets that contain enough materials--administrator's manual, text booklets, questionnaires and answer sheets--to test a classroom), and testing an entire grade within a school at one time, one would be able to reduce administration costs greatly, choose somewhat fewer schools to participate while maintaining similar standard errors, and greatly reduce intrusion in the schools. By offering a program enabling schools to participate in an assessment (see the discussion of "local option" below) one could add substantially to the base of students tested (at the expense of the users, not the project), and offer meaningful results to schools participating in the sample.

---

[1] The sampling frame could be the roster of U.S. primary and secondary schools maintained by the Curriculum Information Center of Market Data Retrieval in Westport, Connecticut. This computerized file is annually updated and contains the range of grades and numbers of students enrolled in public, private and parochial schools, as well as the names of key school district officials.

But there is more at stake than these efficiencies. The utility of a national assessment for state and local users lies, in large part, in providing them with a comparable national sample against which to measure local assessments. By bringing NAEP procedures more in line with local procedures, interpretation would be strengthened considerably. Thus, the move to a simpler scheme of test administration would not significantly reduce the benefits of standardization, and would positively contribute to the comparability of assessments, making NAEP-state comparisons far more valid.

Test Administration

One of the most expensive activities associated with the current approach to NAEP is test administration. The new demands being placed on NAEP make it questionable whether the control derived from existing NAEP procedures is worth the cost or the limitations these procedures place on the assessment, especially when comparability with other assessment programs is considered.

It is the opinion of the HumRRO team that an inordinate amount of current NAEP resources are expended on test administration. We propose major revisions in administration procedures. We believe these proposed changes would create a small, documentable difference in student performance. These changes also would be much less expensive, create less intrusion in sampled schools, and make it more reasonable for states to borrow NAEP administration procedures.

Our central recommendation is that all testing materials be delivered to schools selected for testing and that administration be conducted by local school personnel. We propose a testing plan that would proceed as follows:

1. An initial mailing to the Chief State School Officers, the Council for American Private Education, and the National Association of Independent Schools would alert them to the forthcoming national assessment (including the local option described below), list the schools in the state that have been included in the sample, and request their help in eliciting the cooperation of selected schools.

2. A letter describing the assessment and requesting cooperation from schools selected in the sample to be mailed to the superintendents of the districts containing the selected schools. The superintendent would be encouraged to contact the State Education Agency (SEA) or to call a toll-free number operated by NAEP for further information, if needed. The superintendent would be asked to sign and return a postcard indicating the school's willingness to participate. Follow-up letters and, if necessary, phone calls would be used to obtain a decision from each superintendent.

3. A letter similar to the superintendent's to be sent to the principals of the selected schools. This letter would contain more

detail about administration procedures and would ask the principal to provide the name of the contact person within the school to serve as a coordinator for the assessment. Again, a toll-free number for further information will be provided.

4. The coordinator to be mailed more detailed information about the assessment, and asked to return information about how and when the testing would be conducted within the school. These individuals will be strongly encouraged to use the toll-free number to discuss aspects of the administration unique to their setting or to clarify any obscure points.

5. Approximately one week before the scheduled testing date, testing materials to be delivered to the school directly from Intran. All tests will be packaged in class packs. This convenient form greatly simplifies the distribution of materials within the school. Also included will be a narrated filmstrip that covers the purpose of the testing, the types of results that the school will receive as a result of the testing (similar to the local option reporting described below), and proper administration practices. This filmstrip should be shown to the school's teachers before the test administration date.

6. The tests to be administered to all eligible students at the appropriate grade in the school at one time, with the teachers in the school serving as the administrators. A paced tape that provides administration instructions and serves as a timer would be used to help standardize the administration of the test. (Note that the other functions of the currently-used paced tapes have been eliminated.) The specific date and time would be selected by the school within a range of three weeks specified by NAEP. Testing would take no longer than one class period. At this point, we are uncertain whether to suggest that makeup tests be given to absent students. RMC is currently testing students in the Connecticut Assessment of Educational Progress and requiring makeup testing. Each answer sheet indicates whether a student was tested in the regular session or a makeup session. The results of that testing will strongly influence our opinion on whether makeup testing is needed.

7. In a selected group of schools, a NAEP auditor would be sent to the school on the scheduled date of testing to observe the testing process and to document whether correct procedures have been carried out. The schools would be selected to cover the spectrum of schools participating in the assessment, but with added emphasis placed on sending auditors to sites that might be expected to have a higher than average likelihood of employing inappropriate administration procedures.

8. On a specific date, the materials to be picked up or mailed to Intran for checking and scoring.

9. After scoring, results of the assessment to be provided, upon request, to the schools. These reports would be similar to those provided to schools participating in the local option program (see below).

This approach to administration will provide high quality data at a substantially lower cost than current NAEP practices and will minimize intrusion in the schools. We believe this approach has a number of desirable features to recommend it:

1. It follows protocol. Beginning with the SEA, each person in the "chain of command" is made aware of NAEP and asked for permission to continue the process. This way, districts that require special services can be accommodated. In the Connecticut Assessment of Educational Progress, for example, some of the big city districts requested that RMC channel all materials through them; when the request was followed, they were extremely cooperative.

2. It maintains a clear path of control. Starting with the district superintendent, it requires acknowledgement that the NAEP materials have been received. This way, one can ensure that the materials have been read by the person at one level before contacting the immediately subordinate official. Thus no one in the chain is likely to be surprised and local officials are more likely to be cooperative when contacted about their schools' participation.

3. Questions are readily answered. Maintaining a toll-free "hotline" to answer questions is an important aspect of this process. While every effort should be made to produce materials that are clear, invariably there will be special circumstances and unanswered questions. RMC has used such a service in Connecticut and found that it greatly facilitated the administrative process.

4. Test administration procedures will be clear. The use of carefully field-tested administration manuals, the required viewing of the narrated filmstrip, and the provision of a paced tape for administration will make the administration procedures clear and easy to follow. Complaints about poor administration procedures often are not the fault of the administrators; they frequently are provided with unclear manuals. We believe administrators will do well if properly motivated and equipped with clear and understandable directions.

5. Administrators and students will be motivated. The offer to return test results to the school will serve not only as an effective means for securing the cooperation of the school, but also will encourage administrators and students to take the testing seriously. Since the quality of test data depends upon the motivation of the test taker to do as well as possible, it is important to introduce mechanisms that increase the value of the assessment at the classroom level. If the teacher thinks it is important, the students will also. And they will not be hampered by test anxiety to the extent that they would on a test of individual performance.

We strongly believe in the effectiveness of using local school personnel to administer the tests. It will not only save a great deal of money that can be allocated to other NAEP programs, but also is a procedure that can be duplicated by states, will minimize intrusion in

the schools, will get more educators involved in NAEP (and consequently, aware of it), and will provide assessment results to many schools. We are certain that the approach is feasible; RMC already has implemented major aspects of it in the Connecticut Assessment of Educational Progress and found them to be very effective.

Paced Tapes

In current NAEP procedure, paced audiotapes are used in the administration of tests to many small groups of students drawn from different classes and grades in each participating school. The use of paced audiotapes necessitates a separate examination time for each form of the test, thus complicating the logistics of the assessment.

The use of paced tapes by NAEP is a practice of questionable value when its benefits are weighed against the limitations it creates. Because they were intended to minimize the influence of reading ability on test performance, the paced tapes have had a great deal of appeal for policy makers and the education community in general. The advantages, however, have been largely illusory. The use of tapes does have, according to NAEP studies, a positive effect on the test performance of lower socio-economic groups; however, that effect is quite small--perhaps a few percentage points. While the influence of reading ability on tasks not intended to measure reading skill should be of concern, perhaps it is best handled by careful item development and by separate smaller studies or probes intended to help explain group differences in the major assessments. The continued use of the paced tapes would keep administration costs high, while at the same time placing expensive restrictions on the test package format itself and preventing the use of multiple test packages in a single administration.

The proposed approach for NAEP would eliminate the use of paced tapes except to provide administrative directions. This would allow the administration of multiple forms within a classroom simultaneously. We believe that the advantages of strongly reduced intrusion, and substantially lowered cost, coupled with the minimal impact on results of eliminating paced tapes and the ability to conduct further investigations, justify this procedural revision.

Comparability Issues

In the preceding discussion, we proposed major revisions in NAEP administration procedures. We proposed eliminating trained administrators and paced tapes, changing the age-level samples to grade-level samples, and administering multiple forms of a test simultaneously to an entire grade within a school. We believe that such changes would significantly reduce costs, add substantially to the interpretability of results, and minimize intrusion in the schools. We also believe that,

using the controls and supplemental materials we proposed, such changes would have minimal impact on the results.

Nevertheless, it is necessary to establish mechanisms that will assure the comparability of NAEP results under a new system of administration with previous assessments. Any plan not providing for comparability would make several years of NAEP virtually worthless. Such a constraint does not mean that administration procedures cannot be modified; it simply means that change must be carefully planned, documented, and assessed.

What is required is an orderly transition period during which test performance under the new system is related to performance under the old. One way would be to administer previously administered items to students in a subsample of schools, using an exact duplicate of old NAEP procedures. These schools would also be tested using the new administration procedures, so that any differences exhibited could be analyzed and the impact of the new procedures established.

To be more specific, we suggest that a sample of one-sixth of the schools selected for the full assessment be selected for study. In half these schools, a trained administrator using standard prior NAEP procedures would select samples of 15 students each (one sample for each age group). These samples would be given a booklet of items administered in a previous NAEP assessment. This administration would be carried out before the scheduled time of testing for the full sample of assessment schools, and the schools would be assessed again along with all other selected schools using regular procedures. In the other half of the special school sample, the extra administration using all old NAEP procedures would take place after the regular administration.

During these special administrations, students would receive a package of items from the previous assessment as well as a linking test (these would need to be appropriately counterbalanced.). The linking test, which also would be taken by a nationally representative sample during the regular testing, would be scaled using IRT. The data from such a study would provide an estimate of the change in student performance from the previous assessment, taking into account both any change due to revised administration procedures and true changes in student performance since the previous administration. Then in all future administrations, linkages back to earlier administrations could be straightforward applications of IRT.

While there obviously would be added expense in conducting such a careful study to investigate the consequence of changing administration procedures, the investment would show a healthy return. In future administrations, when labor and travel costs are even higher than they are today, the benefit of having made such a study when it was still financially feasible would be apparent. At some point, current administration practices will become so expensive that they will cripple the assessment effort. Now is the time to plan for the change that must come. The sooner it is done, the less expensive it will be to do, and

the sooner it will free up resources for more valuable aspects of NAEP than test administration.

Local Option

Each state and district in the country could be provided with the opportunity to administer the NAEP assessment to their students at the same time as the selected sample schools. The administration procedures would be identical, although samples of students, rather than whole grades, might be tested at the option of the user. A charge of two dollars ($2.00) per student would cover the expense of printing, delivering, picking up, scanning, scoring and producing reports on the results.

Based upon the experience of RMC in Connecticut, where a local option plan has been offered, we estimate that ten percent of the students at any grade level would be participants in such a program. This would be approximately 300,000 per grade or 900,000 over the three grades to be tested. At a charge of two dollars per student, a local option program could be expected to add $1.8 million to the NAEP operating budget.

A program enabling local participation would provide a wide variety of benefits. Local school districts would focus upon, and be interested in, NAEP results as never before. The objectives, test items, and results would be much more carefully scrutinized by thousands of teachers and local administrators, and the reports produced by NAEP would have more impact. In addition, many districts that cannot afford to develop such an assessment on their own would have a low cost alternative. States, too, might find the local option plan appealing. Connecticut, for example, is paying almost $100,000 to develop social studies tests and to test a sample of approximately 10,000 students. While participating in the NAEP assessment would not supply all the same services that Connecticut is receiving in its $100,000 contract, the availability of such an option might encourage the states to contemplate more specialized assessments of their own, for example, linking various remediation strategies to assessment results.

NAEP would benefit from the local option program as well. With anticipated print runs of only 2,500 per booklet, printing costs per booklet are relatively high for the proposed model. Local option might well add 15,000 copies per booklet to the order. This would drive down costs per booklet considerably. Also, the results from local option would provide additional data for certain analyses. Since the local option districts would be self-selected, their results could not be used to establish normative data, but they would help in developing scales in any item response theory analysis, and in multivariate analyses of the data.

The ability of IRT analyses to facilitate detailed and specific interpretations of test data would, as the approach is gradually established at NAEP, be a strong inducement for schools to take advantage of the local option plan. It can, for example, be used to develop "person-fit" statistics--indicators of how consistently a student responded to the questions. When the person-fit statistics are high for a group of students relative to the school as a whole, it often is an indicator that something improper took place during the administration--something that should be taken into account when interpreting the results. When person-fit statistics are high for most of the students in a school, that often means that the curricular emphases for that school are quite different from the typical pattern. This too can be a signal that interpretation of the test results will warrant special care.

REPORTING, DISSEMINATION, SERVICE AND RESEARCH

The limited impact of NAEP and the limited utility of NAEP data have been of increasing concern in recent years. These shortcomings are a product of a now defunct political necessity as well as the original conceptualization of the National Assessment. The political climate during the early years of NAEP was such that the project had to be operated very conservatively so as to avoid any implication of a Federal take-over of education. While considerable effort could be expended to collect high-quality data, the designers felt that extreme caution had to be exhibited in the presentation of findings. Thus, NAEP's approach to reporting and dissemination has been to report the statistical data, refrain from interpretive discussions, and make the data available to others to analyze and interpret more fully if they have the interest, ability and resources.

Much has been said already about NAEP's "spare-no-expense" approach to data collection procedures. An important result of this orientation, however, is the relatively low expenditures on activities related to reporting, dissemination, service and research. As NAEP's budget was gradually reduced from six to approximately four million dollars, the expensive but firmly entrenched data collection procedures were preserved. ECS's solution to budget cuts has been to do less rather than to alter practices. A prime example is the decision to discontinue assessments in science. Considering the role of science and technology in society generally and the rapidly growing need for personnel in engineering and other technical fields, the decision to drop science assessments was unfortunate in the extreme. As it happened, the National Science Foundation came to the rescue by funding a science assessment which addressed a small subdomain of science education, but that was not foreseen nor can it be depended on in the future.

Dissemination and service, already of relatively less importance in the NAEP scheme, suffered even more from budget cuts. Services provided to states conducting their own assessments, for example, were curtailed. NSF has funded other activities which one might reasonably expect NAEP to finance. It was NSF funds that enabled NAEP to put their user tapes in a reasonable form for researchers to use in secondary analyses and to provide some training in such analyses. NSF also awarded the National Council of Teachers of Mathematics grants to write "interpretive" reports, articles, etc., based on the results of the mathematics assessments. While it will certainly continue to be desirable to seek external funds for such activities, it is also important and possible--both psychometrically and economically--for NAEP to make service, evaluation, and research central to its operation.

Up-Grading User Services

Recognizing the necessity to do more for less, the HumRRO plan includes a substantial reallocation of funds to dissemination and service using monies saved by more economical approaches to test development and data collection.

Much of the discussion in previous sections on development, administration, etc., pertains to matters of utility. This illustrates a very important characteristic of the proposed NAEP design. In the past, ECS has treated the major assessment activities as almost totally independent endeavors. Assessment instruments have been developed to produce data for limited types of analyses which address a limited number of purposes. Clearly, multiple purposes can be accommodated without diminishing the capability for conducting the traditional NAEP analyses. Thus, an improved NAEP will maximize the involvement of NAEP "users" in NAEP operations, including development activities. Researchers, for example, who are interested in secondary analyses of NAEP data would not be faced with the problem of using data not intended to address their research questions if they had been involved in the process from the development of objectives onward.

The continuity of NAEP staff involvement throughout an assessment cycle has already been mentioned. It makes sense that the content experts who developed the instruments participate meaningfully in the writing of reports and in presentations of findings rather than leaving those tasks to technical writers from a publications department who have no expertise in the particular content area and, therefore, are only marginally able to interpret findings or respond to questions. The same logic applies to the people who are targeted as major users--schools, state assessment agencies, researchers, federal agencies, and so on.

Benefits to State and Local Agencies

Our suggestions for increasing the utility of national assessment data are dictated in part by the belief that NAEP has focused too much on what students know or don't know and not enough on possible explanatory factors. We believe that assessment results are meaningful to most practicing educators only when they can relate those results to their own students and to educational programs. Broad generalizations and recommendations, and reporting of large group averages, and conversely, item by item reporting, have little meaning for most educators. As a result, assessment results have little impact. Educators must be able to see their students in the report. In other words, if educators are to be expected to make constructive use of the data, the reports must provide enough descriptions of enough different types of students (the "who") and be able to describe their strengths and weaknesses, explanations of how these came about (the "why"), and how these might be corrected.

NAEP has targeted statewide assessment programs, rather than classroom teachers, as primary users. Given the great complexities of communicating effectively with classroom teachers, this may have been a reasonable choice of primary audience in the early years of the program. But if the future development of the National Assessment is to be in the direction of a heightened emphasis on research and evaluation, then it is also time to address the school and the classroom as well as state assessment. One obvious technical assistance program is the provision of training services to local education officials who wish to use NAEP information. RMC offices in New Hampshire, North Carolina, Missouri, and California are already providing this type of service in connection with Title I evaluations.

It is also time to improve the services offered to the state programs. We believe NAEP has seriously erred in its attempts to be useful to statewide assessment programs. NAEP has taken the basic approach that its obligation was to establish an ideal assessment program and to disseminate information on its practices, materials, and results to SEAs. This approach has failed to recognize two major facts: NAEP operates much too expensively for most SEAs to reproduce their practices; and in the fourteen years that NAEP has operated, many statewide assessment programs have made tremendous progress.

The fact that NAEP is so expensive has created serious difficulties for SEAs wishing to use NAEP data for comparative purposes. Since the states cannot afford to match NAEP's expensive administrative practices, they cannot be confident that their results are comparable to NAEP's. While undoubtedly of high quality, NAEP data often proves to be relatively useless to states, because they cannot afford to collect equivalent data.

The earlier section on test development proposed that NAEP utilize state test items. By using materials provided by states, NAEP would do far more than reduce its item development costs. It would also make its results more usable to states. States would benefit by having their materials used by NAEP. At no cost to them, they would have national difficulty values created for their items, thus giving them the best comparability data available. Of course, the large item pools the new NAEP contractor will be compiling will be of considerable use to state and local officials in their future assessment activities. They could be made available to other qualified users at the cost of reproduction. Rather than just disseminating a few exercises, NAEP could make a substantial pool of items available to the education community. Items of good quality would be included whether NAEP had used them in their assessment or not.

When NAEP began, few states were developing their own testing materials in reading, writing, and mathematics. NAEP was a leader in those areas, and now can reap the benefits of the involvement of many states. It is now time for NAEP to take on a similar leadership role in other content areas. Listening and, to a limited extent, speaking are becoming new areas of focus in basic skills. Facility with computers is

certain to become another in the near future. If NAEP assumes a leadership role in these areas, it will only be a matter of time before the state education agencies begin their own development efforts. When this happens, NAEP could become a central repository of the states' efforts in these areas, just as it could now be in reading, writing, and mathematics.

One of the greatest benefits to local school districts would result from their participation in the local option program, an innovation already working successfully in some state assessment programs. Under this option, a local district can choose to test all of its students (or a representative sample) at a particular grade level at a minimum charge per student. The information reported back to a district could be most useful in evaluating a district's program. Local option reports would not only provide results for the district, but would also include performance data for students in similar types of communities in the larger geographic region and the nation as a whole.

Research

Educational policymakers and practitioners are calling for a national assessment which is more interpretive and provides practical guidance. It is time for NAEP to take responsibility for the kind of secondary analysis activities presently left almost entirely to outside researchers. As mentioned previously, this new role would have an impact on development activities. If NAEP is to be more research oriented, the nature of NAEP test booklets must change since p-values associated with individual items would no longer be the only concern. Purposeful item packaging procedures and instruments measuring a far greater range of variables will be required.

NAEP's service role could be closely tied to its new research emphasis. In addition to servicing educators in the content areas, NAEP could seek out other special interest groups in education, including agencies funding research, to determine what information those groups would like to have which could be gathered through regular assessment mechanisms or by special studies. A mechanism for accomplishing this was described in the development section. In this way, NAEP can address policy-relevant issues such as educational equity, bilingual education, establishing standards of competency, etc. The funds available as a result of markedly reducing development and administration costs could make such activities a regular part of the ongoing National Assessment program.

Thus, considerably more attention would be given to explaining, interpreting and evaluating, and the needs of policymakers would be served to a far greater extent. It seems reasonable to suggest, for example, that the assessment sample include schools that have instituted objectives-based programs, schools that have minimum competency testing programs (with and without accompanying remediation plans), schools with

programs for the exceptionally bright, schools with a special emphasis, such as schools for the performing arts or science high schools, and schools with a variety of organizational structures. Information can be collected on many instructional variables as well, including teaching practices and materials. Not only could changes in these kinds of factors be monitored over time, they could also be related to student achievement using classrooms and schools as the units of analysis. Thus, policy-relevant issues could be addressed via the regular assessment instruments. Smaller separate probes might be appropriate for other issues.

In order to carry out this new program of research and service, the NAEP staff will have to have a strong research capability so that it can take a leadership role in policy analysis and other areas of research as well. Of course, the concerns of various groups of researchers will have been taken into account during assessment development activities so a larger research community will have a greater interest in using NAEP data.

One of the problems associated with secondary analysis of NAEP data in the past has been the difficulty of applying standard techniques to data associated with a complex cluster-sampling design. The proposed administration procedures allow the use of a simpler sampling design as explained in a previous section. This will facilitate secondary analysis considerably. While certain subgroups in the population will still have to be oversampled, individuals can be identified in such a way that if a researcher wants data for a nationally representative random sample, such a sample of cases can be drawn from the NAEP data, still maintaining a relatively large sample size appropriate for various multivariate analysis techniques. Of course, if the researcher wishes to make comparisons of subgroups including those oversampled, then the full NAEP sample can be used. One of the services NAEP could perform is to provide data from a sample tailored to the needs of the individual researcher. The feasibility of some form of automated access to NAEP data might also be investigated--i.e., it might be possible for appropriately cleared researchers to draw a sample from the NAEP data files via direct on-line access.

One idea for simplifying sampling requirements for special studies as well as enhancing reporting and services would be the construction of linking tests for each year's assessments. These linking tests would have great utility not only for the assessment project but for other users as well. At this time, all NAEP materials are administered to random, nationally representative samples of students. When a special analysis is done, another random sample must be drawn, or it must be folded into the already drawn sample. For some studies, it would be far more cost effective to draw representative samples of convenience to whom the linking tests would be administered along with the special study assessments, and then to compare the special study groups to the random samples.

Such linking tests would have great utility to people outside of NAEP as well. Test publishers would find such tests useful for updating and verifying their own norms, as would any district desiring national norms to link into for their own local tests. Because NAEP has a significantly better sample of cooperating schools than any individual concern (such as an independent test publisher), such norms would be highly valued by anyone involved in studies requiring use of normative data.

Reporting and Dissemination

NAEP currently publishes an impressive array of newsletters, studies, and reports of its assessments, which are, of course, the documents of greatest interest. While NAEP assessment reports have been directed to different audiences based on level of technical expertise and level of specificity required, those reports still pertained to the same results. The HumRRO team approach to NAEP would produce far more information addressing a multitude of research questions. Different audiences will be interested in different issues.

We contemplate, therefore, a much expanded publishing list, including the kind of interpretive reports previously left to outside groups and requiring external funding. We also anticipate the production of many unpublished reports, for example, reports to individual schools who have participated in the local option program and reports to state assessment agencies providing national data on items they contributed to or used from the NAEP data pool that were included in a NAEP assessment. Another dissemination activity that would be useful to the states would be publication of lists of states that have used a particular item from the pool so that the states can compare data.

The same teams of NAEP content specialists, research staff, and consultants who have followed each assessment from the development of objectives through test administration will be involved in producing descriptive and interpretive reports of the assessment. This continuity of involvement will promote the highest realizable level of familiarity with and comprehension of an assessment, and will promote a more imaginative understanding of its interpretive possibilities.

In addition to published and unpublished reports, we recommend more frequent participation in workshops, annual meetings of teachers, school administrators, educational researchers, and other professional organizations, and other public meetings that would provide an appropriate forum for the dissemination of information about the National Assessment and the services it offers. A good part of the battle to make NAEP more useful is simply a matter of making it better known.